Collaboration or copying? Student behavior during two-phase exams with individual and team phases

By: Ian D. Beatty

Beatty, I. D. and Harris, L. A. (2015). "Collaboration or copying? Student behavior during two-phase exams with individual and team phases." *Proceedings of the 2015 Physics Education Research Conference* (PERC), College Park MD: July 29-30, 2015 Pages 59-62.

Made available courtesy of American Association of Physics Teachers
http://dx.doi.org/10.1119/perc.2015.pr.010

Abstract:

Students take a two-phase exam twice: once individually, and a second time working in teams. Proponents hope that during the team phase, students will discuss, debate, and resolve questions by sharing their reasoning, challenging each other, and reaching consensus. Potential adopters fear that students might uncritically follow the majority answer or mimic one dominant team member. To explore this empirically, I data-mined students' solo- and team-phase responses from the final exams of three different introductory physics courses to construct multiple measures of team dynamics. My results substantiate prior findings that teams do engage in meaningful debate and explore the virtues of various possible answers. The two-phase exam implementation used does not force teams to submit a common answer, allows students to "hedging their bets" for partial credit, and incentivizes helping teammates.

Keywords: Two-phase exams | assessment | collaborative learning | group dynamics

Article:

I. INTRODUCTION

A *two-phase exam* combines the assessment and accountability virtues of a traditional individual exam with the learning benefits and rapid formative feedback of group problem-solving [1]. First, during the *solo phase*, students work through the exam individually. They then turn in their answer sheets, are assigned to groups for the *team phase*, and collaborate to complete the exam (or a subset or variant of it) a second time. One study in Ocean and Earth Science showed that students of all achievement levels learn more from a two-phase exam than from completing the

same questions twice on their own [2]. Another, in Physics, found only a short-term improvement in subsequent test performance on matched near-transfer questions [3]. A third, also in Physics, found that most students like the two- phase format and believe it helps them learn [4].

Two-phase exams are intended to cause students to articulate their thinking with their teammates, explore each other's' reasoning, resolve differences, and reach a mutually satisfactory consensus answer. Of course, we cannot force students to engage in deep, reasoning-based discussion. Given the high stakes and limited time, one can reasonably fear that some or most teams might defeat our intent by blindly following the team member they perceive as the most knowledgeable, take a vote, or pursue some other mechanical process to choose their team response.

Instructors that implement two-phase exams generally report seeing high levels of student engagement during the team phase, with much visible argumentation—often animated, sometimes heated [1,4]. However, appearances can be deceiving; classroom observations and anecdotal evidence are insufficient to demonstrate that all, or even the vast majority, of teams are really using reasoning-based argumentation to reach their consensus response. Therefore, I have attempted to use students' exam response patterns to infer the degree to which students are genuinely resolving differences by convincing their teammates. This study's guiding research question is, "Does a comparison of students' solo-phase and team-phase responses show any evidence for or against the belief that teams are engaging in substantive discussion?" By "substantive" discussion, I mean team-members seeking to convince or be convinced by their teammates, as opposed to uncritically following one member or adopting the majority response.

II. CONTEXT AND PROCEDURE

The data for this study come from the final exams of three courses that I taught. Courses A (40 students) and B (39 students) were *Introductory Physics I with Calculus*, in Spring 2015 and Spring 2014 respectively. Course C (24 students) was *Introductory Physics II with Calculus* in Fall 2013. These represent three separate populations of students. In each case, the course's final exam was the students' first experience of a two-phase exam.

My implementation of two-phase exams is somewhat unusual, in ways advantageous to this research. Both phases took place during one three-hour final exam block, with 2:00 or 2:15 allocated to the solo phase and the remainder allocated to the team phase. During the solo phase, students recorded their answers on the exam and also on a separate response sheet. At the end of the phase, they turned in the response sheet and kept the exam for the team phase, meaning that they had a record of their solo-phase responses during team discussions. Each student received a new, blank response sheet for the team phase.

The exams were all multiple-choice, with six options per question in Course A, four in Course B, and five in Course C. The exams contained 33, 25, and 24 questions respectively, although not all of these were traditional multiple-choice questions: five, four, and ten respectively were "pick all that apply" (PATA) questions, in which the correct response might require students to select more than one of the options (e.g., "Which of the following objects is/are accelerating?"). In

order to avoid excessive complication and ambiguity during analysis and interpretation, the study reported here analyzes only the traditional "pick one" questions from each exam. Courses A, B, and C contained 28, 21, and 14 of those.

One unusual feature of these exams is that they allowed students to "hedge their bets" on traditional "pick one" (not PATA) questions, opting for the safety of assured partial credit by selecting more than one response option [5]. On each question, students had as many points to allocate as the question had answer options. They could put all those points on one option, earning full credit if correct and zero if incorrect. Alternatively, they could divide their points between two or three options, or allocate one point to every option. They earned the number of points allocated to the correct response, so the more they divided their points, the less they could earn, but the more likely they were to earn some. This feature provides data on how confident students are about their responses, and can reveal when their confidence level changes even though their first-choice answer option might not, which is relevant to the analysis below. It also makes doubt more tangible to students, potentially stimulating discussion.

Courses A, B, and C had 13, 10, and 6 teams respectively. In Course A, I engineered teams of three members (with one foursome), attempting to balance students' midterm exam performances and communication styles. In Courses B and C, I randomly assigned students to teams of four members (with one threesome in Course B). I instructed teams to aspire to a consensus response to each question, but they were free to dissent if they wished. This latter feature is also unusual for two-phase exams; I find that it allays students' fear of giving control of their grade to others, and provides me with data about whether team discussion convinces all members or just a majority— essential to the analysis below.

The third unusual feature of my two-phase exam implementation is the method I use to determine overall exam scores. Rather than using a weighted average of solo- and team-phase scores, I give each student their solo-phase raw score plus a "team bonus." Half of the team bonus comes from the student's solo- to team-phase raw score gain, and half from the average of their teammates' gains. (The actual formula includes some tunable parameters that let me adjust and cap the overall spread and scale of the resulting bonuses, to avoid overly distorting the grades.) I tell students that "If you help your teammates, you win. If your teammates help you, you win. The only way to lose is if nobody improves, or if you convince each other of something incorrect." I find that this scoring system allays student concerns about free riders, and may increase students' investment in convincing or being convinced.

When interpreting what follows, please note that although these exams are largely qualitative and most questions can be answered relatively quickly by someone with sound conceptual understanding, well-structured knowledge, and clear thought, they contain many traps for the unwary or misconception-beset. Even the very best students in the course rarely earn more than 75% of the credit during the solo phase, and therefore appreciate that they have much to gain by revisiting their thinking and seeking input during the team phase.

III. DATA AND ANALYSIS

A. Dissent

A potential indicator of healthy debate among team members is *dissent*. Dissent indicates that not all members were convinced of the wisdom of any one response, which in turn suggests that members demanded to be convinced. Therefore, I examined all questions for which each team had any disagreement during the solo phase (meaning they had something to resolve), and calculated the fraction for which their team-phase responses were not identical to each other: the team's *dissent frequency*. (If a team had non- identical solo responses for ten questions, and identical team-phase responses for seven of those ten, their dissent frequency would be 0.3.)

The distribution of team dissent frequencies for the three courses are summarized in Fig. 1. These strike me as far enough above zero to suggest that members did generally hold their ground unless convinced, but not so high as to indicate an inability or unwillingness to reach consensus.

B. Plurality Rules

One possible undesirable team dynamic is "plurality rules," in which the team simply adopts whichever response had been chosen by the most members during the solo phase. To check for this, I calculated each team's *plurality match frequency*: For all questions on which the team had solo-phase disagreement, but for which a single solo-phase response did have a plurality of support (as opposed to, say, a 2-2 or 1-1-1 tie), and which were resolved to a unanimous team response, I calculated the fraction of cases for which that unanimous response matched the plurality solo-phase choice (i.e., the most popular solo response won).

The distribution of teams' plurality match frequencies is summarized in Fig. 2, showing that that most teams chose their most-popular solo response around 50% of the time. Since the most popular is likely often correct, a frequency this high is not surprising. If teams were following a mechanical "plurality rules" process, however, I would expect values much closer to 100%.

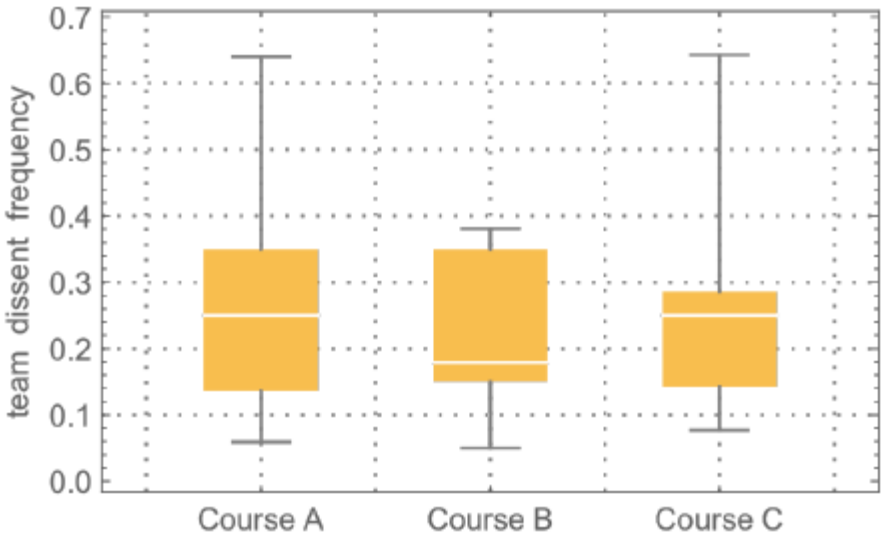

FIG. 1. Distribution of team dissent frequencies. The box represents the 25% to 75% range, with a white line at the median. Whiskers show the lowest and highest values.

C. Leadership

Another possible undesirable team dynamic is "follow the leader," in which the team uncritically mimics one member's solo-phase responses. To tease out the prevalence of this, I calculated a *match frequency* for each team member: the number of times the team's consensus response matched that member's solo-phase response, divided by the total number of resolved questions. I defined the "leader" as the team member with the highest match frequency. (If a team resolves ten questions and their consensus response matches member A's solo response six times, B's five times, and C's two times, their leader match frequency is 0.6.) This is an indicator of team dynamics, because values significantly below one reveal that the team did not follow any one individual particularly faithfully.

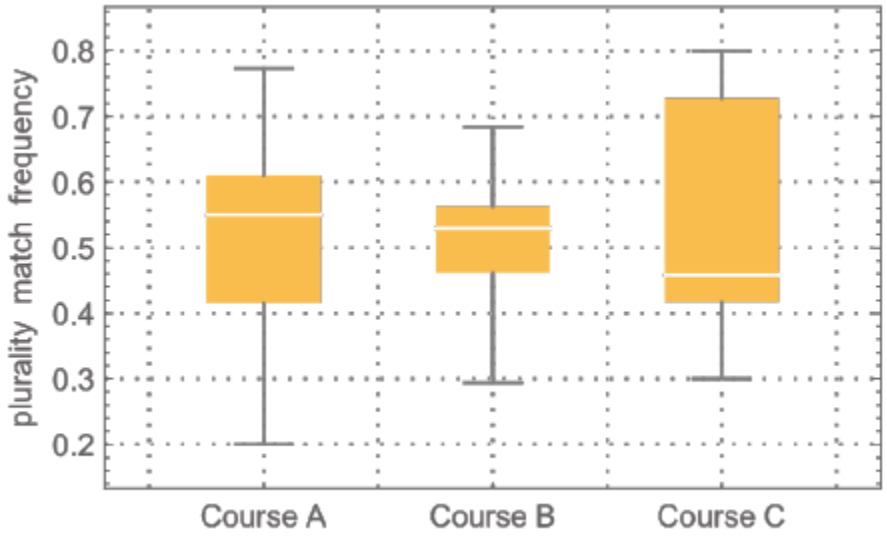

FIG. 2. Distribution of team plurality match frequencies.

Figure 3 shows the distribution of leader match frequencies for each course. We see that every team came to a unanimous decision different from their leader's solo response at least 15-20% of the time, and many teams disagreed with him or her at least half the time. Note that this only counts questions for which consensus was reached, meaning that even the "leader" was convinced to change responses. (Dissent, discussed earlier, also reveals a leader's lack of influence.)

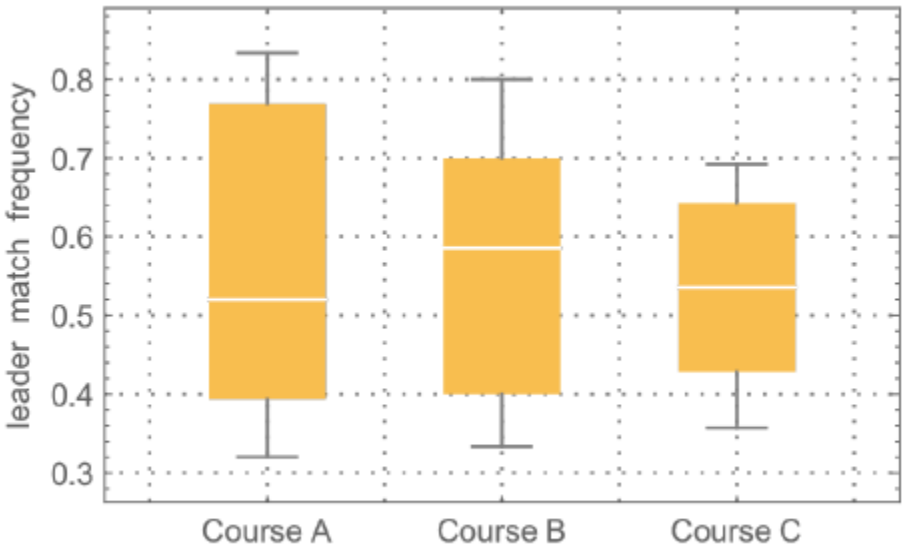

FIG. 3. Distribution of team leaders' match frequencies.

D. Answer-Changing

A different indicator of healthy team dynamics is *answer-changing*. When a team engages in reasoned debate, I expect to see all members changing their minds at least occasionally due to ideas they encounter in the discussion. In follow-the-leader dynamics, some team members may

have large answer-changing counts, but the leader will have a very low count. Thus, the lowest student answer-changing frequency within a team is a measure that suggests debate: The larger this lowest frequency is, the healthier the dynamic. Unlike the previous measures, this one includes all (non-PATA) questions, including those for which the team's solo-phase responses agreed and those they failed to resolve. My logic is that any response- changing between phases suggests that the individual was influenced by team discussion.

Figure 4 shows the distribution of this measure for the teams in each course. The figure reveals that all students changed at least some of their responses, and that in the vast majority of teams, even the most confident or stubborn member did so often enough to suggest a true give-and-take between members. (For comparison, the distribution of change frequencies for *all* students—not just the least- changing from each team—has a median value of 0.48, 0.52, or 0.60 for the three courses.)

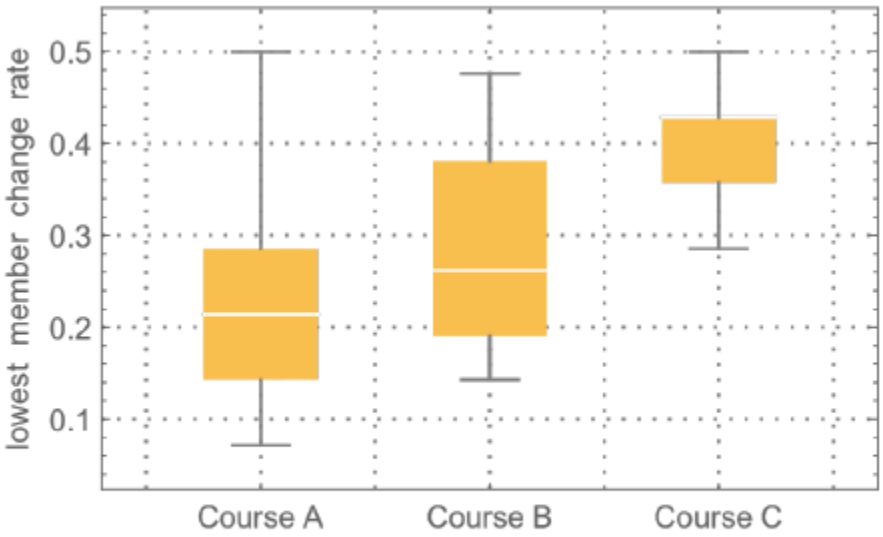

FIG. 4. Distribution of answer-changing rates, taking the lowest rate from each team. (For Course C, the median rate is equal to the 75th percentile at 0.43.)

E. From Nowhere

Yet one more indicator of healthy team dynamics is the occasional selection of a *from-nowhere* response: a consensus response that no member had chosen during the solo phase. I expect this to be relatively rare, but any occurrence indicates that the team's discussions go beyond having each member defend his/her choice, and are rich enough to occasionally discover that all members had erred. Figure 5 shows a stacked histogram of the number of from- nowhere responses each team had. We see that for all courses combined, a majority of teams (17 of 29) discovered at least one.

IV. CONCLUSIONS & DISCUSSION

Taken together, the five measures of team dynamics presented above suggest that during the two-phase final exam in each of the three courses studied, most or all teams engaged in

substantive argumentation and chose consensus answers on their merits rather than according to which or how many members proposed them. Teams' *dissent frequencies* were high enough to indicate that failing to reach consensus was a very real option for most teams, which in turn suggests that members did not yield their positions unless convinced. *Plurality match frequencies* were consistent with the expectation that the most common solo-phase response would often win out, but was not chosen automatically or uncritically. *Leader match frequencies* showed that no member's solo-phase responses were adopted uniformly, and that most teams did not follow any one member particularly closely. Similarly, *answer- changing frequencies* revealed that in almost all teams, all members changed their responses a significant fraction of the time due to team deliberation. Finally, the fact that a majority of teams came to consensus at least once on an answer that no member had initially chosen—*from-nowhere responses*—shows that these teams' discussions were adequate to discover a perspective not initially championed by any member.

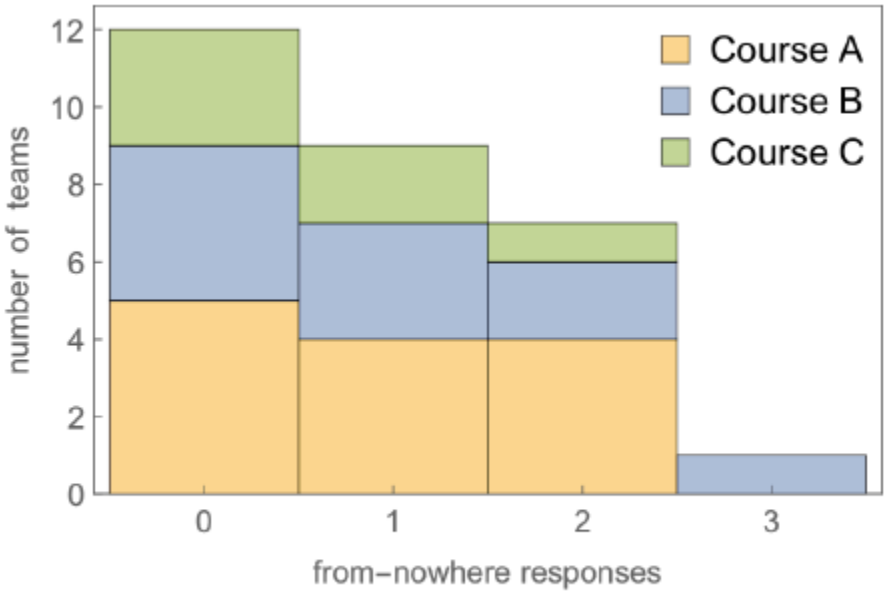

FIG. 5. Frequency of teams reaching consensus responses not chosen by any member during the solo phase.

These findings corroborate and substantiate my observations as the exam invigilator: that all teams worked through the questions systematically, resolving disagreements by asking each member to justify his or her choice. Debates were common, even heated, and I frequently saw weaker students challenging stronger ones until they understood the argument put forward.

This analysis does not reveal whether *all* team members were fully engaged in the discussion and resolution. It is possible, perhaps likely, that some teams have one free- rider member who listens without contributing while the other two or three members debate. I explored some potential indicators of this, but did not find any convincing.

Another interesting question is whether team dynamics correlates with team performance. Using team members' solo- to team-phase normalized gain [6] to eliminate the strong anticorrelation of

gain with solo-phase score, I explored several possible correlations with the above measures of team dynamics, but found nothing convincing.

While richer insight could be gained through systematic observation and/or video-analysis of group interactions, the approach presented here is far less labor-intensive and can easily and economically be repeated in other two-phase exam contexts. Future research could profitably examine how differences in two-phase exam implementation affect team dynamics.

[1] C.E. Wieman, G.W. Rieger and C.E. Heiner, Phys. Teach. 52, 51 (2014). DOI: 10.1119/1.4849159

[2] B. Gilley and B. Clarkston, J. Coll. Sci. Teach. 43, 3 (2014). [3] J. Ives, in *Proceedings of the 2014 Physics Education Research Conference* (2015). DOI: 10.1119/perc.2014.pr.027

[3] J. Ives, in *Proceedings of the 2014 Physics Education Research Conference* (2015). DOI: 10.1119/perc.2014.pr.027

[4] G. W. Rieger and C. E. Heiner, J. Coll. Sci. Teach. 43, 4 (2014).

[5] L.K. Michaelsen, A.B. Knight and L.D. Fink, *Team-based Learning: A Transformative Use of Small Groups* (Praeger, 2002), ISBN 9780897898638, p. 230.

[6] R.R. Hake, Am. J. Phys. **66**, 1 (1998). DOI: 10.1119/1.18809